OUTLIERS: Feb. 4, 2005, R. Dudley, 18.465 notes

A rough definition of an outlier is that it's an observation far away from the bulk of the data. There may be multiple outliers in a given data set, especially if it's large. For example, Bill Gates's wealth would be an outlier among those of all individuals.

One of the main ideas of robustness is to use statistical procedures that are not sensitive to outliers. If there are outliers in a data set $X_1, ..., X_n$, then at least one of the extreme order statistics $X_{(1)}$ or $X_{(n)}$ must be an outlier.

The ordinary sample mean $\overline{X}$ and sample variance $s_X^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2$ are not robust because $\overline{X}$ and, even more, s_X^2 are sensitive to outliers. If just one large observation $X_{(n)}$ becomes arbitrarily large, then both $\overline{X}$ and s_X^2 go to $+\infty$.

On the other hand, nonparametric methods based on ranks, such as the Wilcoxon rank-sum test and the runs test, are not at all sensitive to outliers. If $X_{(n)}$ is made larger it still keeps the same rank, n, so the values of the nonparametric statistics don't change at all. The same is true if $X_{(1)} \to -\infty$, when it keeps the same rank 1.

But, what is an outlier? It turns out to be even harder to give a precise definition than for a sample quantile. Some books give examples of outliers and a few try to give specific rules for identifying them.

An example, given in a book by D. Freedman, Pisani, and R. Purves, is that an observation 5 standard deviations away from the mean would be an outlier. For a normal distribution, the probability of such an observation is less than $6 \cdot 10^{-7}$, so from a truly normal distribution, such a thing wouldn't happen except very rarely or in a large data set. Put another way, normal distributions don't tend to produce outliers: as n gets large, $X_{(n)}$ tends to grow, but only slowly, of the order of $\sqrt{\log n}$, so $X_{(n)}$ won't be much larger than $X_{(n-1)}$, and so on. Or, if one thinks one has a normal distribution but gets an observation 5 standard deviations from the mean, in a sample with not too large n, say $n < 10,000$, that observation must not really be from the normal distribution, it must be from some other distribution, sometimes called a contaminating distribution. It might have resulted from some error, or a wrong normality assumption.

There's a problem though with defining outliers in terms of standard deviations if the standard deviation is estimated from the sample, because the sample standard deviation is itself so much influenced by the outlier. Specifically, just looking at the formula for sample variance, to have an observation X_j five or more sample standard deviations away from the sample mean, $|X_j - \overline{X}| \geq 5s_X$, requires a sample size easily seen to be at least 26 (since 26-1=25) and in fact at least 27. But we can recognize outliers in smaller samples than that. If you take Bill Gates's wealth, together with that of 9 other people chosen at random from the population to form a sample of size 10, you will see that Gates's wealth is an outlier by the rough definition.

Other people have tried to define outliers precisely as follows. Define the lower quartile of the sample q_1 as the 1/4 quantile and the upper quartile q_3 as the 3/4 quantile (recalling however that for samples, quantiles have slightly varying definitions). The *interquartile range* IQR is defined as $q_3 - q_1$. That's a scale statistic that's robust, not sensitive to outliers: if we move data in the lower quarter or upper quarter of the order statistics outward, it won't change the IQR. An attempted definition of outlier is an observation that's distant by at least $3IQR$ from the interval $[q_1, q_3]$.

But here's an example where that definition doesn't work well. Let X_j be observations on amount of precipitation (rain, or water equivalent of snow) per day over a year. Suppose that on at least 3/4 of all days in the year, there is no measurable rain or snow at a given location (maybe, a relatively dry one, but not all that dry). Then q_1 and q_3 will both be 0, so $IQR = 0$. So by the attempted definition, any precipitation at all would be called an outlier, which doesn't seem right.

If there was precipitation on more than 1/4 of all days, but less than half, it could be that q_3, although positive, is quite small and so IQR, which equals q_3 in this case, is small. So we'd be calling amounts of rain "outliers" if they were larger than $4q_3$ which might still not be that large.

By the way, the median rainfall per day in either case would be 0, which is very uninformative about rainfall.

It seems that we might only want to call a daily amount of rain or snow an outlier if we compared it to for example the 10 or 20 days with most rain in a typical year. So the choice of what to call an outlier may depend on what kind of data we're looking at, not on any universal numerical rule.